INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Mixture of linear mixed models Application to repeated data clustering

Gilles Celeux, Christian Lavergne, Olivier Martin

N° 4566

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--4566--FR+ENG

INRIA
RHÔNE-ALPES

Mixture of linear mixed models
Application to repeated data clustering

Gilles Celeux, Christian Lavergne, Olivier Martin

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet is2

Rapport de recherche n° 4566 — — 20 pages

Abstract: The problem of finite mixture analysis from repeated data is considered. Data variability is taken into account through linear mixed models leading to a mixture of mixed models. The maximum likelihood estimation of this family of models through the EM algorithm is presented. The problem of selecting a particular mixture of mixed models is considered. Illustrative Monte Carlo experiments are presented and an application to the clustering of gene expression profiles is detailed. All those experiments highlight the interest of linear mixed model mixtures for taking account of data variability in a proper way.

Key-words: Random Effect; Linear Model; Mixture Model; Cluster Analysis; Gene Expression Profile; Microarray data.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 Montbonnot-St-Martin (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Mélange de modèles linéaires mixtes
Application à la classification de données répétées

Résumé : Nous proposons un modèle de mélange pour des données répétées. Ce modèle prend en compte la variabilité des données par des modèles linéaires mixtes associés à chaque composant du mélange. Nous présentons l'estimation des paramètres de ce modèle par la méthode du maximum de vraisemblance via l'algorithme EM. Nous étudions aussi le problème de sélection d'un modèle particulier. Des expérimentations de Monte-Carlo et une application à un problème de classification de profils d'expression de gènes sont présentées. Elles illustrent la capacité de notre modèle à prendre en compte efficacement la variabilité des données.

Mots-clés : Effets aléatoires, modèle linéaire, modèle de mélange, classification, profils d'expression de gènes, puces à ADN.

1 Introduction

Finite Mixture analysis is a powerful tool of data analysis. In particular multivariate Gaussian mixtures have been proved useful in the modelling of heterogeneity in a cluster analysis context (see [13]). However, finite mixture models are not tailored for taking account of variability of random effects with a potentially infinite number of levels. Two common attitudes when facing such a variability problem for a mixture model are the following

- Neglecting the problem by basing the mixture model estimation from a single measure of the variables for each statistical unit.
- Restricting the variability to a mean effect by estimating the mixture model from the mean values of R independent repeated measures of each statistical unit.

Both attitudes can be expected to be unsatisfactory. The first one clearly jeopardizes the model as soon as the variability is important. The second one is assuming that the variability does not depend on covariates or on the statistical units and can be unrealistic.

In this article, we propose taking account of the variability of measurements, in the model-based clustering context, by embedding a linear model with random effects in a mixture distribution. In Section 2, the general framework of linear mixed models is presented and is illustrated with some models of interest for the application presented in Section 4.2. In Section 3 the mixture of mixed models and its estimation through the EM algorithm is presented, and the EM equations are detailed for some models experimented in Section 4. Section 4 is devoted to the presentation of illustrative numerical Monte Carlo experiments and to an application in the context of gene expression data clustering, a problem which motivated the present research. A short concluding section summarizes the main points of this article.

2 Linear mixed models

Linear mixed models, hereafter abbreviated L2M, are aiming to analyze the variability that is evident in data by including both fixed and random effects (see [16]). The L2M we considered here obeys the equation

$$\mathbf{y} = \underbrace{X\beta}_{\text{fixed effects part}} + \underbrace{U\xi}_{\text{random effects part}} + \epsilon \tag{2.1}$$

where

- $\mathbf{y}$ is the random vector of N observations,
- $X_{(N,p)}$ and $U_{(N,q)}$ are known design matrices,
- β is the fixed effect vector, of size p, to be estimated,

- $\xi = (\xi_1', \ldots, \xi_H')'$ is the vector of the H random effects, with ξ_h of size q_h for $h = 1, \ldots, H$, such that the ξ_h's are independent, $\xi_h \sim \mathcal{N}(0, \tau_h^2 \mathrm{Id}_{q_h})$, for $h = 1, \ldots, H$, and the variances τ_h^2, $h = 1, \ldots, H$ are to be estimated,
- ϵ is a random vector of residuals of size N such that $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_N)$ with σ^2 to be estimated and ϵ is independent of each ξ_h.

Thus $\mathbf{y}$ is a Gaussian vector $\mathcal{N}(X\beta, \sigma^2 \mathrm{Id}_N + U\Gamma U^t)$ where Γ is the $q \times q$ diagonal matrix

$$\Gamma = \begin{bmatrix} \tau_1^2 \mathrm{Id}_{q_1} & & \\ & \ddots & \\ & & \tau_H^2 \mathrm{Id}_{q_H} \end{bmatrix}, \text{ with } \sum_{h=1}^{H} q_h = q.$$

To simplify the presentation of the mixture of L2M models that we propose in this article, we restrict attention to the linear mixed models that we considered in the following in a cluster analysis purpose. But, there is no difficulty to extend this L2M mixture model to any linear mixed model of the form (2.1).

Let assume that R repetitions of measures are recorded at T different times for I independent statistical units. Denoting y_{itr} the r^e repetition of the measure at time t for statistical unit i, the following L2M, taking into account some different covariance structures in the repetitions, are possible

$$\begin{aligned} \text{(E1)} &: y_{itr} = \beta_t + \xi_i + \epsilon_{itr} && (2.2) \\ \text{(E2)} &: y_{itr} = \beta_t + \xi_{it} + \epsilon_{itr} && (2.3) \\ \text{(E3)} &: y_{itr} = \beta_t + (\xi_i + \xi_{it}) + \epsilon_{itr} && (2.4) \end{aligned}$$

where

- β_t represents the fixed effect of time,
- $\xi_i \sim \mathcal{N}(0, \omega^2)$ is the random effect of the unit i on the measure,
- $\xi_{it} \sim \mathcal{N}(0, \tau^2)$ is the random effect of the unit i at time t,
- $\epsilon_{itr} \sim \mathcal{N}(0, \sigma^2)$ is the error measure.

Those three models differ by the covariance structure they involve between two measurements for a statistical unit. Noting $\delta_k^{k'} = 1$ if $k = k'$ and 0 otherwise, those covariance structures are as follows.

Model E1 : $y_{itr} = \beta_t + \xi_i + \epsilon_{itr}$

$$\mathrm{cov}(y_{itr}, y_{i't'r'}) = \omega^2 \delta_i^{i'} + \sigma^2 \delta_i^{i'} \delta_t^{t'} \delta_r^{r'}. \quad (2.5)$$

For a given statistical unit, the covariance between two measures is equal to ω^2, independently of time and repetition. The variance of an observation is $\omega^2 + \sigma^2$.

Model E2 : $y_{itr} = \beta_t + \xi_{it} + \epsilon_{itr}$

$$\mathrm{cov}(y_{itr}, y_{i't'r'}) = \tau^2 \delta_i^{i'} \delta_t^{t'} + \sigma^2 \delta_i^{i'} \delta_t^{t'} \delta_r^{r'}. \tag{2.6}$$

In this model, the covariance between the repetitions for the same statistical unit and for the same time is not null and is equal to τ^2. The variance of an observation is $\tau^2 + \sigma^2$.

Model E3 : $y_{itr} = \beta_t + \xi_i + \xi_{it} + \epsilon_{itr}$

$$\mathrm{cov}(y_{itr}, y_{i't'r'}) = \omega^2 \delta_i^{i'} + \tau^2 \delta_i^{i'} \delta_t^{t'} + \sigma^2 \delta_i^{i'} \delta_t^{t'} \delta_r^{r'} \tag{2.7}$$

This model is the most complex one which will be considered here. The covariance between two different times for the same statistical unit is equal to ω^2. At a fixed time, the covariance between the repetitions of a statistical unit is $\omega^2 + \tau^2$. The variance of an observation is $\omega^2 + \tau^2 + \sigma^2$.

In this article, for simplicity, the presentation is focused on model E2. Actually, this model is often realistic and leads to simple interpretation. But models E1 and E3 can also be of interest.

The canonical equation (2.1) applies for $E2$ with

- $\mathbf{y} = (\mathbf{y}_1{}', \ldots, \mathbf{y}_I{}')'$ where $\mathbf{y}_i$ is the random vector of measures for unit i of size RT,
- $\xi = (\xi_{11}, \ldots, \xi_{1T}, \ldots, \xi_{I1}, \ldots, \xi_{IT})$ random effect vector of size IT and $\xi \sim \mathcal{N}(0, \tau^2 \mathrm{Id}_{IT})$,
- the design matrix

$$U_{(N,IT)} = \begin{bmatrix} 1_R & 0_R & \cdots & 0_R \\ 0_R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0_R \\ 0_R & \cdots & 0_R & 1_R \end{bmatrix} \tag{2.8}$$

  where 1_R and 0_R denotes respectively the vectors $(1, \ldots, 1)'$ and $(0, \ldots, 0)'$ of size R,
- ϵ: random vector of errors of size N and $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_N)$,
- β: unknown fixed effects vector of size T,
- and the design matrix $X_{(N,T)}$ as follows

$$X_{(N,T)} = \begin{bmatrix} \dot{X} \\ \vdots \\ \dot{X} \end{bmatrix} \tag{2.9}$$

where

$$\dot{X}_{(RT,T)} = \begin{bmatrix} 1_R & 0_R & \cdots & 0_R \\ 0_R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0_R \\ 0_R & \cdots & 0_R & 1_R \end{bmatrix}. \tag{2.10}$$

L2M are incomplete data models where the missing data are the random effects $\xi_{it}, i = 1, \ldots, I; t = 1, \ldots, T$. The maximum likelihood parameter estimates can be derived for instance with the EM algorithm [12] which consists of maximizing iteratively the conditional expectation, knowing the observations and a current value of the parameters, of the complete likelihood

$$\begin{aligned} l(\beta, \tau^2, \sigma^2 | \mathbf{y}, \xi) &= -\frac{1}{2}(N + IT)\ln(2\pi) - \frac{1}{2}N\ln(\sigma^2) - \frac{1}{2}IT\ln(\tau^2) \\ &\quad -\frac{1}{2}\frac{(\mathbf{y} - X\beta - U\xi)'(\mathbf{y} - X\beta - U\xi)}{\sigma^2} - \frac{1}{2}\frac{\xi'\xi}{\tau^2}. \end{aligned} \tag{2.11}$$

Detailed formulas for the EM algorithm for mixed models can be found for instance in [16], Section 8.3 or in [19].

3 Mixture of linear mixed models

In this section, the finite mixture model is extended to the L2M context in order to propose a model-based cluster analysis tool for repeated data. Since Gaussian mixture is the most employed mixture model especially in a cluster analysis context, we restrict attention to this model. In a multivariate Gaussian mixture model, it is assumed that an observation $\mathbf{y}$ is arising from the mixture distribution

$$f(\mathbf{y}) = \sum_{k=1}^{K} p_k \phi(\mathbf{y} | \mu_k, \Sigma_k) \tag{3.12}$$

where $p_k \geq 0, k = 1, \ldots, K$ are the mixing proportions verifying $\sum_{k=1}^{K} p_k = 1$, $\phi(.|\mu_k, \Sigma_k)$ being the density of a Gaussian distribution with mean vector μ_k and variance matrix Σ_k. Consequently, knowing the mixture component C_k from which an observations arises, its conditional distribution is a Gaussian distribution with mean μ_k and variance matrix Σ_k.

3.1 Mixture model for repeated data

To take account of repeated data in the mixture framework, we simply add the assumption that the repeated measures of a statistical unit belong to the same mixture component. This natural assumption allows us to embed L2M in the mixture framework. Then, the specific L2M assumptions to be added to the mixture model (3.12) concern the component mean μ_k and variance matrix Σ_k. It is assumed that $\mathbf{y}^k$, the vector of observations arising from mixture component C_k, obeys a L2M equation of the form

$$\mathbf{y}^k = X^k\beta_k + U^k\xi^k + \epsilon^k \tag{3.13}$$

β_k being the fixed effect vector, ξ^k the random effect vector, X^k and U^k the design matrices.

To clarify the presentation, we now detail those specific assumptions for the model E2. The observations are arising from one of the K components and those that come from component C_k define a random vector $\mathbf{y}^k$ of size $N_k = I_kTR$ where I_k is the number of statistical units belonging to C_k. The vectors $\mathbf{y}^k$, $k = 1, \ldots, K$, verify the equation

$$y^k_{itr} = \beta_{kt} + \xi^k_{it} + \epsilon^k_{itr}. \tag{3.14}$$

Thus

$$\mathbf{y}^k \sim \mathcal{N}_{N_k}(X^k\beta_k, \tau^2_k U^k(U^k)' + \sigma^2 \mathrm{Id}_{N_k}) \tag{3.15}$$

where

- $X^k_{(N_k,T)}$ is a design matrix with the same structure as the design matrix X defined in (2.9) and (2.10),
- β_k is the fixed effect vector of size T for component C_k, $\beta_k = (\beta_{kt}, t = 1, \ldots, T)$,
- $U^k_{(N_k,I_kT)}$ is a design matrix with the same structure as the design matrix U defined in (2.8),
- τ^2_k is the variance of the random effect for component C_k,
- σ^2 is the residual variance which is assumed here independent of the mixture components.

Furthermore, the random effect variables $\xi^k_i = (\xi^k_{it}, t = 1, \ldots, T)$ being a vector of size T and ϵ^k are assumed independent.

Remark:
We have considered above a mixture model where parameters β_k et τ^2_k are dependent from k, while σ^2 is fixed over the K components. Obviously, many alternative models can be considered. Possible models of interest included

- M1: $(\beta_k, k = 1, \ldots, K; \tau^2, \sigma^2)$,

- M2: $(\beta_k, \tau_k^2, k = 1, \ldots, K; \sigma^2)$, namely the above mentioned model,
- M3: $(\beta_k, \tau_k^2, \sigma_k^2, k = 1, \ldots, K)$.

In order to estimate the parameters of a L2M mixture, we consider the maximum likelihood approach. We make use of the EM methodology that takes into account the incomplete structure of the data. Here missing data are of two types: *(i)* the indicator vectors $\mathbf{z} = (\mathbf{z}_i,\ i = 1, \ldots, I)$ of the statistic unit memberships to the mixture components: $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ with $z_{ik} = 1$ if $i \in \mathcal{C}_k$ and 0 otherwise, *(ii)* the random effects $\xi_i^k,\ i = 1, \ldots, I$, for each component.

3.2 EM algorithm for L2M mixture

We detail the EM algorithm for the L2M mixture model E2-M2, (namely a mixture of E2-type L2M, with different fixed and random effects for each mixture component and a same residual variance), which seems to be a model of great interest. We denote $p = (p_1, \ldots, p_K)$ the mixture proportions, $(\theta_k = \beta_k, \tau_k^2, \sigma^2)$ the parameters of the linear mixed model associated to component C_k and $\theta = (\beta_1, \ldots, \beta_K, \tau_1^2, \ldots, \tau_K^2, \sigma^2)$. As noted above, there are two types of missing data: the indicator vectors $\mathbf{z} = (\mathbf{z}_i, i = 1, \ldots, I)$ and the random effects $(\xi_{it}^k, t = 1, \ldots, T)$, for each unit i in $\mathcal{C}_k$.

The log-likelihood associated to the complete data $(\mathbf{y}, \mathbf{z})$ for model E2-M2 is

$$l(\theta, p|\mathbf{y}, \mathbf{z}, \xi) = \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \ln \left(p_k f_k(\mathbf{y_i}, \xi_i^k | \theta_k) \right) \tag{3.16}$$

where the vector $\mathbf{y_i}$ of size TR contains all the measured values for unit i and where $\xi_i^k = (\xi_{i1}^k, \ldots, \xi_{iT}^k)$ denotes the random effect vector of unit i in $\mathcal{C}_k$. Conditionally on the component C_k from which it arises, vector $\mathbf{y_i}$ is a realization from a $\mathcal{N}(\dot{X}\beta_k, \Gamma_k)$ with $\Gamma_k = \tau_k^2 \dot{U}\dot{U}' + \sigma^2 \mathrm{Id}_{TR}$ where

$\dot{U}_{(TR,T)} = \begin{bmatrix} 1_R & 0_R & \cdots & 0_R \\ 0_R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0_R \\ 0_R & \cdots & 0_R & 1_R \end{bmatrix}$ is the design matrix associated to $\mathbf{y_i}$.

Therefore, f_k is the density of Gaussian distribution with mean $\mu_k = \begin{bmatrix} \dot{X}\beta_k \\ 0_T \end{bmatrix}$ and variance matrix $\Sigma_k = \begin{bmatrix} \Gamma_k & \tau_k^2 \dot{U} \\ \tau_k^2 \dot{U}' & \tau_k^2 \mathrm{Id}_T \end{bmatrix}$.

Thus, we have

$$\begin{aligned} l(\theta, p|\mathbf{y}, \xi, \mathbf{z}) &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln(p_k) + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln(f_k(\mathbf{y_i}, \xi_i^k|\theta_k)) \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln(p_k) + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}h(\theta_k|\mathbf{y_i}, \xi_i^k)) \end{aligned} \tag{3.17}$$

where

$$\begin{aligned} h(\theta_k|\mathbf{y_i}, \xi_i^k) &= -\frac{1}{2}(TR+T)\ln(2\pi) - \frac{1}{2}TR\ln(\sigma^2) - \frac{1}{2}T\ln(\tau_k^2) \\ &\quad -\frac{1}{2}\frac{(\mathbf{y_i} - \dot{X}\beta_k - \dot{U}\xi_i^k)'(\mathbf{y_i} - \dot{X}\beta_k - \dot{U}\xi_i^k)}{\sigma^2} \\ &\quad -\frac{1}{2}\frac{(\xi_i^k)'(\xi_i^k)}{\tau_k^2}. \end{aligned} \tag{3.18}$$

E step

At iteration $q > 0$, this step consists of computing the expectation of the complete log-likelihood knowing the observed data and the current value of the parameter $\theta^{[q]}, p^{[q]}$, $[q]$ representing the iteration index. In the L2M mixture context it is

$$\begin{aligned} \mathcal{Q}(\theta, p|\theta^{[q]}, p^{[q]}) &= \mathbb{E}\left(l(\theta, p|\mathbf{y}, \mathbf{z}, \xi)|\mathbf{y}, \theta^{[q]}, p^{[q]}\right) \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} t_i^{[q]}(k)\ln(p_k) \\ &\quad + \sum_{i=1}^{I}\sum_{k=1}^{K} t_i^{[q]}(k)\mathbb{E}[h(\theta_k|\mathbf{y_i}, \xi_i^k)|\mathbf{y}, \theta^{[q]}] \end{aligned} \tag{3.19}$$

where

$$t_i^{[q]}(k) = P(i \in \mathcal{C}_k|\mathbf{y_i}, \theta^{[q]}, p^{[q]}) = \frac{p_k^{[q]} f_k(\mathbf{y_i}|\theta_k^{[q]})}{\sum_{l=1}^{K} p_l^{[q]} f_l(\mathbf{y_i}|\theta_l^{[q]})} \tag{3.20}$$

denotes the conditional probability that $\mathbf{y}_i$ arises from component C_k.

Since $(\xi_i^k)'(\xi_i^k)$ and $(\mathbf{y_i} - \dot{U}\xi_i^k)$ are sufficient statistics for the complete model (see for instance [16]), there is no need to compute the expectation $\mathbb{E}\left(l(\theta, p|\mathbf{y}, \mathbf{z}, \xi)|\mathbf{y}, \theta^{[q]}, p^{[q]}\right)$. To proceed to the maximization of $\mathcal{Q}(\theta, p|\theta^{[q]}, p^{[q]})$, in the M step, we only need to compute the conditional expectation of those sufficient statistics $(\xi_i^k)'(\xi_i^k)$ and $(\mathbf{y_i} - \dot{U}\xi_i^k)$, knowing observed data $\mathbf{y_i}$ and a current value for the parameters $\theta^{[q]}$ and $p^{[q]}$ (cf. [3]).

It is done easily because, knowing that $\mathbf{y}_i \in C_k$, following Trottier [19] p. 49, we get

$$\begin{aligned}\mathbb{E}(\xi_i^{k'}\,\xi_i^k|\mathbf{y_i},\theta^{[q]}) &= \tau_k^4(\mathbf{y_i}-\dot{X}\beta_k)'\Gamma_k^{-1}\dot{U}\dot{U}'\Gamma_k^{-1}(\mathbf{y_i}-\dot{X}\beta_k)\\ &\quad +R\tau_k^2-\tau_k^4\mathrm{tr}(\Gamma_k^{-1}\dot{U}\dot{U}')\end{aligned} \tag{3.21}$$

and

$$\mathbb{E}(\mathbf{y_i}-\dot{U}\xi_i^k|\mathbf{y_i},\theta^{[q]}) = \dot{X}\beta_k+\sigma^2\Gamma_k^{-1}(\mathbf{y}_i-\dot{X}\beta_k). \tag{3.22}$$

M step

This stage consists of finding the values maximizing $\mathcal{Q}(\theta,\pi|\theta^{[q]},\pi^{[q]})$. From (3.19), it leads to

$$p_k^{[q+1]} = \sum_{i=1}^{I}\frac{t_i^{[q]}(k)}{I},\ \text{for }, k=1,\ldots,K, \tag{3.23}$$

and to solve the following log-likelihood equations for parameters β_k, τ_k^2, $k=1,\ldots,K$ and parameter σ^2

$$\sum_{i=1}^{I} t_i^{[q]}(k)\frac{\partial\mathbb{E}[h(\theta_k|\mathbf{y_i},\xi_i^k)|\mathbf{y},\theta^{[q]}]}{\partial\beta_k}=0, \tag{3.24}$$

$$\sum_{i=1}^{I} t_i^{[q]}(k)\frac{\partial\mathbb{E}[h(\theta_k|\mathbf{y_i},\xi_i^k)|\mathbf{y},\theta^{[q]}]}{\partial\tau_k^2}=0, \tag{3.25}$$

and

$$\sum_{i=1}^{I}\sum_{k=1}^{K} t_i^{[q]}(k)\frac{\partial\mathbb{E}[h(\theta_k|\mathbf{y_i},\xi_i^k)|\mathbf{y},\theta^{[q]}]}{\partial\sigma^2}=0. \tag{3.26}$$

Using the conditional expectations of the sufficient statistics (18) and (19), it leads to the following explicit formulas:

$$\begin{aligned}\sigma^{2[q+1]} &= \frac{1}{N}\sum_{i=1}^{I}\sum_{k=1}^{K} t_i^{[q]}(k)\left[\sigma^{4[q]}(\mathbf{y_i}-\dot{X}\beta_k^{[q]})'\ \Gamma_k^{-1[q]}\Gamma_k^{-1[q]}(\mathbf{y_i}-\dot{X}\beta_k^{[q]})\right.\\ &\quad \left.+RT\sigma^{2[q]}-\sigma^{4[q]}\mathrm{tr}(\Gamma_k^{-1[q]})\right],\end{aligned} \tag{3.27}$$

$$\begin{aligned}\tau_k^{2[q+1]} &= \frac{1}{T\sum_{i=1}^{I} t_i^{[q]}(k)}\sum_{i=1}^{I} t_i^{[q]}(k)\left[\tau_k^{4[q]}(\mathbf{y_i}-\dot{X}\beta_k^{[q]})'\ \Gamma_k^{-1[q]}\dot{U}\dot{U}'\Gamma_k^{-1[q]}(\mathbf{y_i}-\dot{X}\beta_k^{[q]})\right.\\ &\quad \left.+T\tau_k^{2[q]}-\tau_k^{4[q]}\mathrm{tr}(\Gamma_k^{-1[q]}\dot{U}\dot{U}')\right],\ \text{for } k=1,\ldots,K,\end{aligned} \tag{3.28}$$

$$\beta_k^{[q+1]} = \frac{1}{\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \left[\sigma^{2[q]} (\dot{X}' \dot{X})^{-1} \dot{X}' \Gamma_k^{-1[q]} (\mathbf{y_i} - \dot{X} \beta_k^{[q]}) + \beta_k^{[q]} \right], 1 \leq k \leq K. \tag{3.29}$$

Remarks:

- In this paper, we focused for simplicity on model E2-M2. The generalization to alternative models is straightforward. For instance, we give hereunder the M step formulas for the model E2-M3 that we experiment in Section 4. In the M step, parameters $(\beta_k, \tau_k^2, \sigma_k^2)$ are updated as follows for $k = 1, \ldots, K$

$$\begin{aligned} \sigma_k^{2[q+1]} &= \frac{1}{TR \sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \left[\sigma^{4[q]} (\mathbf{y_i} - \dot{X} \beta_k^{[q]})' \ \Gamma_k^{-1[q]} \Gamma_k^{-1[q]} (\mathbf{y_i} - \dot{X} \beta_k^{[q]}) \right. \\ &\left. + RT \sigma_k^{2[q]} - \sigma_k^{4[q]} \mathrm{tr}(\Gamma_k^{-1[q]}) \right], \end{aligned} \tag{3.30}$$

$$\begin{aligned} \tau_k^{2[q+1]} &= \frac{1}{T \sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \left[\tau_k^{4[q]} (\mathbf{y_i} - \dot{X} \beta_k^{[q]})' \ \Gamma_k^{-1[q]} \dot{U} \dot{U}' \Gamma_k^{-1[q]} (\mathbf{y_i} - \dot{X} \beta_k^{[q]}) \right. \\ &\left. + T \tau_k^{2[q]} - \tau_k^{4[q]} \mathrm{tr}(\Gamma_k^{-1[q]} \dot{U} \dot{U}') \right], \end{aligned} \tag{3.31}$$

$$\beta_k^{[q+1]} = \frac{1}{\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \left[\sigma_k^{2[q]} (\dot{X}' \dot{X})^{-1} \dot{X}' \Gamma_k^{-1[q]} (\mathbf{y_i} - \dot{X} \beta_k^{[q]}) + \beta_k^{[q]} \right]. \tag{3.32}$$

- We have presented the EM algorithm for a L2M mixture model using the ML method for estimating the vector parameter θ. An alternative method for estimating the parameters of an L2M model is the restricted maximum likelihood (REML) method, which can be regarded as method of estimation the variance components by maximizing the marginal likelihood obtained by integrating the likelihood over the fixed effect parameter β. In the mixture context, we do not consider REML estimation. Actually, it seems that there is no sensitive difference between EM and REML estimates in L2M (see [16], Section 6.7) and considering REML estimation in this context would involve technical complications without providing the ML estimation of the fixed effect parameters β_k, $k = 1, \ldots, K$. This is a real drawback because maximum likelihood value enters the composition of useful criteria as BIC to select a model (see Section 4.2). Thus, we think that there is little to gain by considering REML estimation in the L2M mixture context.

4 Numerical Experiments

In this section, we report the results of numerical experiments on both simulated and real data sets. Simulation experiments are aiming to assess the ability of the EM algorithm to

correctly estimate L2M mixture parameters. Experiments on a real data set is aiming to highlight the interest of L2M mixture model for clustering repeated data.

4.1 Monte Carlo experiments

For each Monte Carlo experiment, we generated 100 samples from each type of simulated data. Two E2-M2 mixture models, denoted (A) and (B) in the following, have been simulated. In both cases, I is fixed to 200, the number T of instants was three and the number R of repetitions is four and a three component M2 mixture is considered. The mixing proportions were $p_1 = 0.3, p_2 = 0.5$ and $p_3 = 0.2$. Fixed effect parameters were $\beta_1 = (0, 0, 2)^{'}$, $\beta_2 = (-1, 0, -1)^{'}$ and $\beta_3 = (1, 2, 0)^{'}$. The random effect variances were $\tau_1^2 = 0.2, \tau_2^2 = 0.5$ and $\tau_3^2 = 1$. Models (A) and (B) only differ by the error measure variance: For model (A), it is $\sigma_{(A)}^2 = 2$ and $\sigma_{(B)}^2 = 3$ for model (B).

Table 1 displays the mean and into parentheses the standard error of the estimate parameters obtained with the EM algorithm described in Section 3.2. For all the data sets, the EM algorithm has been initiated from an hierarchical clustering designed with the Ward criterion [21]. Table 2 provides the correct classification rate using the maximum a posteriori (MAP) decision rule from the estimate parameter values $\hat{p}, \hat{\theta}$ obtained with EM. This decision rule consists of assigning all the measures of statistical unit i to the mixture component $k(i)$ such that $k(i) = \arg\max_k \widehat{t_i(k)}$ where $\widehat{t_i(k)} = P(i \in \mathcal{C}_k | \mathbf{y_i}, \widehat{p}, \widehat{\theta})$.

Table 1 shows that both models provide sensible estimation of the model parameters. As expected, the estimation accuracy depends on the random effect variances: the greater the variance is, the greater the estimation standard error is. In the same manner, Table 2 shows that the error classification rate increases with the random effect variances. And, as σ^2 increases(model (B)), the imprecision of the parameter estimates increases even if the mean estimates remain good. In the same way, by comparing the results in Table 2 for models (A) and (B), we note that the correct classification rate decreases when σ^2 increases.

To assess the importance of repetitions for estimating L2M mixture models, we carry out additional Monte Carlo experiments. They consist of 100 replications of model (A) but the number of repetitions is $R = 2$ instead of $R = 4$. We denoted (A') this occurrence of model (A). The results for model (A') in Table 2 clearly show that the correct classification rate increases with the number of repetitions. This confirm us in the opinion that it is important to take account properly of repetitions to get relevant clustering structures for highly variable data sets.

4.2 Application to microarray data

Microarrays [4] are one of the latest breakthroughs in experimental molecular biology, which offer the ability to measure the expression levels of thousands of genes simultaneously. Results from multiple experiments are represented by an expression matrix in which each column represents a single microarray experiment and each row represents the expression vector for a particular gene. Clustering of gene expression profiles is used to find co-regulated and

	Parameter	Model	Component 1	Component 2	Component 3
(1)	fixed effects $t=1$		$\beta_1^1=0$	$\beta_1^2=-1$	$\beta_1^3=1$
(2)		(A): $\sigma^2=2$	0.029 (0.124)	-1.007 (0.125)	1.071 (0.274)
(2)		(B): $\sigma^2=3$	0.019 (0.141)	-1.044 (0.156)	1.035 (0.420)
(1)	fixed effects $t=2$		$\beta_2^1=0$	$\beta_2^2=0$	$\beta_2^3=2$
(2)		(A): $\sigma^2=2$	-0.008 (0.131)	0.004 (0.131)	2.061 (0.373)
(2)		(B): $\sigma^2=3$	-0.009 (0.163)	-0.014 (0.163)	2.030 (0.401)
(1)	fixed effects $t=3$		$\beta_3^1=2$	$\beta_3^2=-1$	$\beta_3^3=0$
(2)		(A): $\sigma^2=2$	1.994 (0.153)	-0.992 (0.123)	0.008 (0.256)
(2)		(B): $\sigma^2=3$	1.970 (0.208)	-0.999 (0.165)	-0.037 (0.330)
(1)	proportions		$p_1=0.3$	$p_2=0.5$	$p_3=0.2$
(2)		(A): $\sigma^2=2$	0.301 (0.030)	0.501 (0.041)	0.197 (0.045)
(2)		(B): $\sigma^2=3$	0.308 (0.043)	0.487 (0.060)	0.204 (0.069)
(1)	random effects		$\tau_1^2=0.2$	$\tau_2^2=0.5$	$\tau_3^2=1$
(2)		(A): $\sigma^2=2$	0.211 (0.092)	0.484 (0.117)	0.901 (0.270)
(2)		(B): $\sigma^2=3$	0.216 (0.134)	0.449 (0.153)	0.866 (0.303)
	error measure				
(2)		(A): $\sigma^2=2$		2.005 (0.069)	
(2)		(B): $\sigma^2=3$		2.995 (0.095)	

(1) Simulated parameter values.
(2) Mean and (standard error) for parameter estimations.

Table 1: Parameter estimation values for EM from 100 simulated models (A) and (B).

Simulation	Cluster 1	Cluster 2	Cluster 3
model (A)	91.70	93.62	76.35
model (B)	88.43	88.84	71.12
model (A')	85.38	86.18	65.45

(A) Correct classification rate with $R=4$ repetitions and $\sigma^2=2$.
(B) Correct classification rate with $R=4$ repetitions and $\sigma^2=3$.
(A') Correct classification rate with $R=2$ repetitions and $\sigma^2=2$.

Table 2: Correct classification rates from 100 simulations for models (A), (B) and (A').

functionally related groups. Among the most used methods, we can cite hierarchical classification [6], self-organizing maps [17] and the K-means method [18]. More recently, Yeund *et al.* [23] and Gosh *et al.* [9] based their approach on multivariate Gaussian mixture.

But, Lee *et al.* [10] show that any single microarray experiment is subject to substantial variability and that replicates provide a more reliable analysis of gene expression. However, previous clustering studies of genes expression profiles did not consider this variability prob-

lem and did not use the replicate information. The methodology we present was especially developed for this problem: Owing to the important number of genes, gene effect can be regarded as a random effect. We first give a technical presentation of the considered data set.

4.2.1 The data

After an injury, rapid regeneration of the epithelium is critical to maintain barrier function. The complex process of wound healing involves several steps, including spreading of the cells at the wound edge, migration, and eventually proliferation of the surrounding cells. This process may be disturbed by various factors, such as free radicals, proteases, inflammatory cytokines, which prolong the inflammatory state, and perpetuate a "non-healing" state.

We used the cDNA microarray technology to profile wound healing after a mechanical epithelial injury. The wound was performed using an helicoidal scarificator. The data sets we considered concern the study of human keratinocytes treated with SB203580 drug, a specific inhibitor of the p38 MAP kinase activity. This application concerns 999 genes (or 999 probes on the chips more exactly). The aim is to analyze expression profiles for a cinetic of $T = 4$ times (30 min., 90 min., 180 min. and 360 min. after the cells have been wounded). For each time of the cinetic, the principle of the experiment consists of incorporating the Cy5 fluorophor in the test sample of mRNA (wounded and treated) and the Cy3 in the reference sample (no wounded and no treated). Moreover, for each experiment, duplications are realized in order to obtain two measures for a given gene. These duplications are obtained by designing two spots on the glass with the same probe. So, for each gene, we got $R = 2$ repetitions for the couple of measures (Cy5, Cy3). In the next paragraph, we detail the pretreatment we made.

Pretreatment Before clustering gene expression profiles, two stages appeared to be necessary.

- *Data normalization*
  To delete some variability sources due to technical and biological problems, data are normalized following the approach proposed by Yang *et al.* [22]. This normalization allows a correction depending on intensity level of the spot and is different for each print-tip used to set down the probes on the glass.

- *Detecting differentially expressed genes*
  For detecting differentially expressed genes, we opt for the approach of Newton [14]. This empirical Bayes approach presents the interest of providing the posterior probability that a gene is differentially expressed. More precisely, along a line described in [11], a score function is designed for each gene in $[-1, 1]$. A negative score means that the gene is down regulated, while a positive score means that it is induced. Thus, we get an easily interpretable and standardized score avoiding the drawbacks of measuring differential expression with the gene expression ratios. Actually, the computation

of those ratios needs a standardization exaggerating gene values with small variations in time. This problem has been discussed by several authors (including [20, 5]) when using the Student statistic to identify differentially expressed genes.

A gene is declared differentially expressed if for at least one measure at any time it is declared differentially expressed according to the Newton procedure [14]. For this data set $I = 58$ genes out of 999 are declared differentially expressed. The L2M mixture models were performed on those $I = 58$ genes since the 941 remaining genes are of no interest for the gene expression profiles study.

4.2.2 The cluster analysis

In cluster analysis, the user is facing two problems: He has to choose a clustering criterion and to decide how many clusters they are. One of the advantages of model-based cluster analysis is that it provides principled statistical approaches to compare different solutions. It gives systematic guidance for choosing a relevant model and number of clusters. For instance some of the most employed criteria for choosing the number of clusters have been conceived in the finite mixture context (see for instance [8]). One of the most performing criteria for choosing a mixture model is the BIC (Bayesian Information Criterion) (see for instance [7], [15]). BIC criterion has been defined in a non informative Bayesian framework to approximate the integrated likelihood of a model. For a model M, BIC (a criterion to be minimized) is the opposite of the maximum of the log-likelihood for model M plus $(\nu_M/2)\ln(n)$, ν_M being the number of free parameters in model M, and n being the sample size. On another performing criterion in the cluster analysis context is the ICL (Integrated Completed Likelihood) criterion which is an *à la* BIC approximation of the integrated complete likelihood (see [2] or [13], Sections 6.10 and 6.11).

For data at hand, we compare the performances of 63 different L2M mixtures: The nine models Ei-Mj, for $i = 1, 2, 3$ and $j = 1, 2, 3$ have been considered with $K = 2$ to $K = 8$ components. Table 3 displays the BIC and ICL values for the model E2-M3. It can be noticed that for all the other models, BIC and ICL values were greater than 240. Thus, both criteria strongly support model E2-M3. From this table, it appears that both criteria indicate that the $K = 4$ component model can be preferred and that the $K = 5$ solution can also be interesting. The parameter estimations for $K = 4$ are given in Table 4. We observe that the third class contains only the two highly up regulated genes included in the data set by biologists in order to control the experiment. Thus, we apply all the models after discarding these two genes. It appeared that ICL and BIC criteria proposed on this modified data set a E2-M3 model with three components identical to the previous components of the four component solution. Consequently, we conclude that the E2-M3 classification we selected is stable and robust against outliers.

On Figure 1, the four clusters obtained from the MAP operator for model E2-M3 are depicted according to a representation suggested by Eisen [6]. It makes appear the up regulated genes in red and the down regulated in green. Moreover, for each cluster all the repetitions of the genes are represented.

	2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters
BIC	251.45	214.19	189.41	192.90	206.70	201.50	242.83
ICL	243.82	208.15	183.13	187.42	202.53	199.12	241.64

Table 3: BIC and ICL criteria for model E2-M3

	Parameter β	Cluster proportion	random effect standard deviation	error measure
Cluster 1	$\beta_1 = (0.073, 0.321, 0.416, 0.347)$	$p_1 = 0.367$	$\tau_1 = 0.078$	$\sigma = 0.018$
Cluster 2	$\beta_2 = (-0.057, -0.19, -0.313, -0.517)$	$p_2 = 0.195$	$\tau_2 = 0.066$	$\sigma = 0.004$
Cluster 3	$\beta_3 = (0.763, 0.928, 0.997, 0.955)$	$p_3 = 0.034$	$\tau_3 = 0.01$	$\sigma = 0.000$
Cluster 4	$\beta_4 = (-0.122, -0.023, -0.156, -0.075)$	$p_4 = 0.403$	$\tau_4 = 0.053$	$\sigma = 0.063$

Table 4: Parameters estimations for clustering.

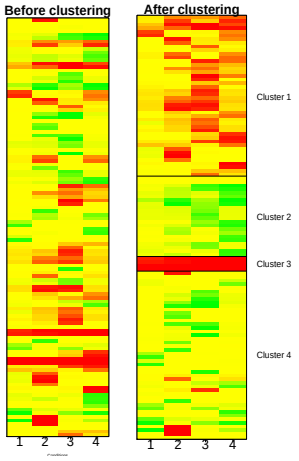

Figure 1: Expression profiles clustering.

With results of Table 4, this figure deserves two types of remarks. The first one concerns the biological interpretation and the second one the measurements quality and the similarity of genes expression profiles for a cluster.

- Clusters 1 and 2 are in opposition especially at third and fourth instants. Cluster 1 is characteristic for up regulated genes whereas genes in cluster 2 are increasingly down regulated with time.

Cluster 3 contains only the two highly up regulated genes included in the date set by biologists in order to control the experiment.

Cluster 4 contains the genes whose expression does not seem different between the two experimental conditions.

- Clusters 1 and 2 have larger random effect standard deviation. This suggests that for those clusters the random effect is more significant. The Eisen representation shows that there are differences between these cluster members. But standard deviation error is smaller than random effect. We conclude that in these clusters, we have good measurements reproducibility but some differences appear between the gene expression profiles. (The clusters dispersion is not explained by variability measurements but by different profiles of genes.)

  Cluster 4 has the largest standard error deviation. This expresses a bad reproducibility of measurements. Consequently, it is difficult to interpret this cluster from a biological point of view.

Finally, it is of interest to see what happen if the variability is restricted to a mean effect neglecting the possible random effects. The following Table 5 provides the BIC and ICL values for this mixture model we called E0-M3. It is remarkable that both criteria highly decrease as K increases. It shows that this model is irrelevant for this data set and could lead to unreliable clustering of the genes since it does not take account of differences in the *gene* variability.

	2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters
BIC	415.11	401.59	349.01	328.59	355.95	317.14	285.12
ICL	414.38	400.65	348.14	326.45	353.01	315.37	284.22

Table 5: BIC and ICL criteria for model E0-M3

5 Discussion

A mixture of L2M models has been proposed and estimated with the EM algorithm. It can be useful in situations where repeated measures are available. In a cluster analysis context, it is expected to lead to more reliable clustering structures since it allows to take profit of the powerful L2M methodology in the mixture framework. And, in many situations, it could be crucial to distinguish the statistical units according to their variability.

In microarray data analysis it could have many applications and could become a reference method for clustering gene expression profiles when the variability is important. Moreover, using the BLUP (Best Linear Unbiased Predictor), see [16], can be useful to build realistic profiles, providing a precise representation of each gene it its cluster.

In Psychometry, it could be also a quite useful tool for cluster analysis since most of the data sets in psychometry are efficiently analyzed with random effects models.

As we have seen, the L2M mixture model can lead to numerous models which can be powerful in specific situations. This is the reason why, it is important to propose a reliable way to select an honest model. In this paper, we have proposed to choose a L2M mixture with the BIC criterion, or the ICL criterion when cluster analysis is the main concern. First experiments show encouraging results for those criteria. Notice that, in the application on microarray data, AIC criterion [1] has systematically selected 8 components. This tendency of AIC to overestimate the number of components has been confirmed on simulated data.

Moreover, the interpretation of L2M mixture models can lead to subtle interpretations, useful for the practitioners since they are sophisticated models without needing many parameters. (For instance mixture components can have the same fixed effects but different random effects, etc)

6 Acknowledgments

We wish to thank Drs Pascal Barbry, Gilles Ponzio and Manal A. Dayem, of the Physiology Genomic Laboratory (Institute of Molecula and cellular Pharmacology, Sophia Antipolis, France) to have provided us their experimental data on the study of the human keratinocytes cicatrization. We also thank Dr. Xavier Gidrol and the members of the Service de Génomique Fonctionnelle (ECA, SGF, Evry, France) for several discussions on the DNA chips technology.

References

[1] H. Akaike. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] C. Biernacki, G. Celeux, and G Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE, Trans. on PAMI*, 22:719–725, 2000.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.

[4] D.J. Duggan, M. Bitnner, Y. Chen, P. Meltzer, and J.M. Trent. Expression profiling using cdna microarrays. *Nature Genetics Supplement*, 21:10–14, 1999.

[5] B. Efron, R. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment, 2001. Dept Statistics, Stanford Univ, May 2001.

[6] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95:14863–14868, 1998.

[7] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 41:578–588, 1998.

[8] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[9] D. Ghosh and A.M. Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18:275–286, 2002.

[10] M.L.T. Lee, F.C. Kuo, G.A. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences*, 97:9834–9839, 2000.

[11] O. Martin. Puces à ADN et analyse de l'expression des gènes. In *XXXIII Journées de Statistique*, pages 566–570, 2001.

[12] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, 1997.

[13] G. McLachlan and D. Peel. *Finite mixture models*. Wiley, 2000.

[14] M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsu. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8:37–52, 2001.

[15] K. Roeder and L. Wasserman. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92:894–902, 1997.

[16] S.R. Searle, G. Casella, and C.E. Mc Culloch. *Variance components*. John Wiley and Sons, 1992.

[17] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expresion with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96:2907–2912, 1999.

[18] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.

[19] C. Trottier. *Estimation dans les modèles linéaires généralisés à effets aléatoires*. PhD thesis, Institut National de Polytechnique de Grenoble, 1998.

[20] V.G. Tusher, R. Tibshirani, and T. Hastie. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98:5116–5121, 2001.

[21] J.H. Ward. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58:234–244, 1963.

[22] Y.H. Yang, S. Dudoit, P. Luu, and T Speed. Normalization for cDNA microarray data. In M.L. Bittner, Y. Chen, A.N. Dorsel, and E.R. Dougherty, editors, *Microarrays: Optical Technologies and Informatics, Proceedings of SPIE*, volume 4266, 2001.

[23] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987, 2001.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399